Learning and Representing Topic

A Hierarchical Mixture Model for Word Occurrences in Document Databases

Thomas Hofmann
Center for Biological and Computational Learning, MIT
Cambridge, MA 02139, USA, hofmann@ai.mit.edu

Abstract

This paper presents a novel statistical mixture model for natural language learning in information retrieval. The described learning architecture is based on word occurrence statistics and extracts hierarchical relations between groups of documents as well as an abstractive organization of keywords. To train the model we derive a generalized, annealed version of the Expectation–Maximization (EM) algorithm for maximum likelihood estimation. The benefits of the model for interactive information retrieval and automated cluster summarization are experimentally investigated.

1 Introduction

Intelligent processing of text and documents ultimately has to be considered as a problem of natural language understanding. In this paper, I will present a statistical approach to learning of language models for context–dependent word occurrences and discuss the applicability of this model for interactive information retrieval. The proposed technique is purely data–driven and does not make use of domain–dependent background information, nor does it rely on predefined document categories or a given list of topics. The presented cluster–abstraction model (CAM) is a statistical mixture model [6, 5] which organizes groups of documents in a hierarchy. Compared to most state-of-the-art techniques based on agglomerative clustering (e.g., [4, 1, 9]) is has several advantages and additional features As a *generative model* the most important advantages are: (i) a sound foundation on the likelihood principle (likelihood as a global clustering criterion), (ii) the probabilistic inference mechanism, (iii) evaluation of generalization performance for model complexity control, (iv) efficient model fitting by the EM algorithm, (v) explicit representation of conditional independence relations. Additional advantages are provided by the hierarchical nature of the model, namely: (vi) multiple resolution levels of document clustering, (vii) discriminative topic descriptors for document groups, (viii) coarse-to-fine approach by deterministic annealing.

2 Probabilistic Clustering of Documents

Let us emphasis the clustering aspect by first introducing a simplified, non–hierarchical version of the CAM which performs 'flat' probabilistic clustering and is closely related to the model proposed in [7] for word clustering. Let the symbols d_i ($1 \le i \le N$) and w_j ($1 \le j \le M$) denote documents and words (word stems), respectively. Counts for word w_j in document d_i are denoted by n_{ij} and $n_i = \sum_j n_{ij}$ is the total number of words in document d_i. Following the standard mixture approach, it is assumed that each document belongs to one out of K clusters C_α. These *hidden variables* are represented by indicator functions $H_{i\alpha} \in \{0, 1\}$, i.e., $H_{i\alpha} = 1$ if d_i belongs to C_α. Moreover, let us introduce parameters $p_{j|\alpha} = P(w_j|C_\alpha)$ for cluster-specific word probability distributions. Then we can specify a joint probability model by[1],

$$P(d_i, H_{i\alpha} = 1|p, \pi) = \pi_\alpha \prod_{j=1}^{M} (p_{j|\alpha})^{n_{ij}} , \tag{1}$$

where π_α are parameters for the prior distribution of $H_{i\alpha}$. The factorial expression for the joint probability reflects conditional independence assumptions about word occurrences (bag-of-words model). Starting from (1) the standard EM approach [2] yields the following coupled re-estimation equations

$$P(H_{i\alpha} = 1|p, \pi) = \frac{\pi_\alpha \prod_{j=1}^{N} (p_{j|\alpha})^{n_{ij}}}{\sum_{\nu=1}^{K} \pi_\nu \prod_{j=1}^{N} (p_{j|\nu})^{n_{ij}}}, \tag{2}$$

$$\pi_\alpha^{new} = \frac{1}{N} \sum_{i=1}^{N} P(H_{i\alpha} = 1|p^{old}, \pi^{old}), \qquad p_{j|\alpha}^{new} = \frac{\sum_{i=1}^{N} P(H_{i\alpha} = 1|p^{old}, \pi^{old})\, n_{ij}}{\sum_{i=1}^{N} P(H_{i\alpha} = 1|p^{old}, \pi^{old})\, n_i} . \tag{3}$$

These equations are very intuitive: The posteriors encode a probabilistic clustering of documents, while the conditionals $p_{\bullet|\alpha}$ represent *average* word distribution for documents belonging to group C_α. Of course, the simplified flat clustering model defined by (1) has several deficits. Most severe are the lack of a multi-resolution structure and the inadequacy of the 'prototypical' distributions $p_{\bullet|\alpha}$ to emphasis discriminative or characteristic words (they are in fact dominated by the most frequent word occurrences). To cure this flaws is the task of the hierarchical extension presented in the next section.

3 Document Hierarchies and Abstraction

Most hierarchical document clustering techniques utilize agglomerative algorithms which generate a cluster hierarchy or dendogram as a by–product of successive cluster merging. In the CAM we will use an *explicit abstraction model* instead to represent hierarchical relations between document groups. This is achieved by extending the 'horizontal' mixture model of the previous section with a 'vertical' component that captures the specificity of a particular word w_j in the context of a document d_i. It is thus assumed that each word occurrence (d_i, w_j) was generated from an *abstraction level* $\mathcal{A}_\nu$, where abstraction levels are identified with inner or terminal nodes of the cluster hierarchy (cf. Figure 1 (a)).

To formalize the sketched ideas, additional hidden variables $V_{(i,j)\nu} \in \{0, 1\}$ are introduced for each word occurrence with $V_{(i,j)\nu} = 1$ if (d_i, w_j) was generated from $\mathcal{A}_\nu$. The hidden variables have to fulfill the following sets of constraints: $\sum_\alpha \sum_{\mathcal{A}_\nu \uparrow \mathcal{C}_\alpha} H_{i\alpha} V_{(i,j)\nu} = 1$, where $\mathcal{A}_\nu \uparrow \mathcal{C}_\alpha$ denotes the nodes $\mathcal{A}_\nu$ 'above' $\mathcal{C}_\alpha$, i.e., nodes on the path to $\mathcal{C}_\alpha$. A pictorial representation can be found in Figure 1 (b): if d_i is assigned to $\mathcal{C}_\alpha$ the choices for abstraction levels of occurrences are restricted to the 'active' (highlighted) vertical path.

Generalizing the non–hierarchical model, a probability distribution $p_{\bullet|\nu}$ over words is attached to each node (inner or terminal) of the hierarchy. The complete data model is given by

$$P((d_i, w_j), H_{i\alpha} = 1, V_{(i,j)\nu} = 1|p, \pi, \rho) = \pi_\alpha \rho_{\nu|(i,\alpha)} p_{j|\nu}, \tag{4}$$

where the additional ρ parameters are prior probabilities for V ($\rho_{\nu|(i,\alpha)} = 0$ whenever $\mathcal{A}_\nu \not\uparrow \mathcal{C}_\alpha$ in the given tree). The prior probabilities $\rho_{\nu|(i,\alpha)}$ capture document–specific distribution over abstraction levels (conditioned on the fact that d_i belongs to $\mathcal{C}_\alpha$). Marginalization over the hidden variables results

[1]For simplicity the number of words in a document is not treated as a random variable and assumed to be given a priori.

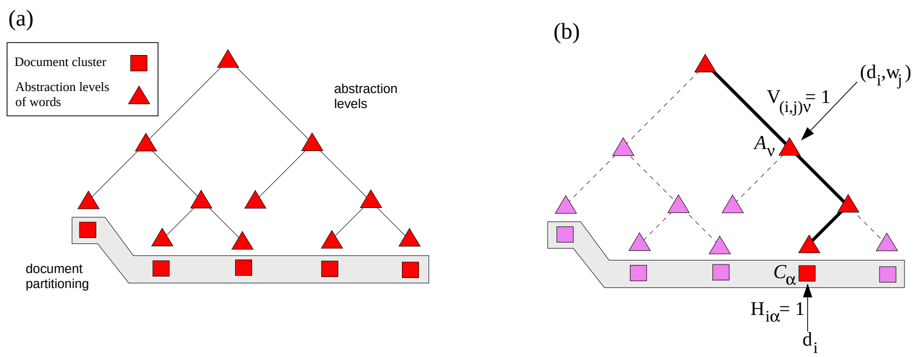

Figure 1: (a) Sketch of the cluster–abstraction structure, (b) the corresponding representation for assigning occurrences to abstraction levels in terms of hidden variables.

in the following log–likelihood of the 'double' mixture model

$$\mathcal{L} = \sum_{i=1}^{N} \log \sum_{\alpha=1}^{K} \pi_\alpha \prod_{j=1}^{M} \left(\sum_\nu \rho_{\nu|(i,\alpha)} p_{j|\nu} \right)^{n_{ij}} . \tag{5}$$

As for the simplified model before, we will derive an EM algorithm for model fitting. The E–step requires to compute (joint) posterior probabilities of the form $P(H_{i\alpha} V_{(i,j)\nu} = 1|p^{old}, \pi^{old}, \rho^{old})$ (abbreviated by $q^{ij}_{\alpha\nu}$ in the sequel). After decomposing by the chain rule one obtains

$$P(V_{(i,j)\nu}{=}\, 1|H_{i\alpha}{=}\, 1, p, \rho) = \frac{\rho_{\nu|(i,\alpha)} p_{j|\nu}}{\sum_\mu \rho_{\mu|(i,\alpha)} p_{j|\mu}}, \quad P(H_{i\alpha}{=}\, 1|p, \pi) \propto \pi_\alpha \prod_{j=1}^{M} \left(\sum_\nu \rho_{\nu|(i,\alpha)} p_{j|\nu} \right)^{n_{ij}} . \tag{6}$$

The M–step re-estimation equations are given by

$$p^{new}_{j|\nu} = \frac{\sum_{i=1}^{N} \sum_{\alpha=1}^{K} q^{ij}_{\alpha\nu}\, n_{ij}}{\sum_{i=1}^{N} \sum_{\alpha=1}^{K} q^{ij}_{\alpha\nu}\, n_i}, \quad \rho^{new}_{\nu|(i,\alpha)} \propto \sum_{j=1}^{M} q^{ij}_{\alpha\nu}\, n_{ij}, \quad \pi^{new}_\alpha \propto \sum_{i=1}^{N} P(H_{i\alpha} = 1|p^{old}, \pi^{old}). \tag{7}$$

Finally, it may be worth taking a closer look at the predictive word probability distribution $p_{j|i}$ in the CAM which is given by $p_{j|i} = \sum_\alpha P(H_{i\alpha} = 1|p, \pi) \sum_\nu \rho_{\nu|(i,\alpha)} p_{j|\nu}$. If we assume for simplicity that $P(H_{i\alpha} = 1|p, \pi) = 1$ for some $\alpha = \alpha_0$, then the word probability of d_i is modeled as a mixture of occurrences from different abstraction levels $\mathcal{A}_\nu$. This reflects the reasonable assumption that each document contains a certain mixture of words ranging from general terms of ordinary language to highly specific technical terms and specialty words.

There are three important problems which need also to be addressed in a successful application of the CAM: First, one has to avoid the problem of *overfitting*. Second, it is necessary to specify a method to determine a meaningful tree topology including the maximum number of terminal nodes. And third, one may also want to find ways to reduce the sensitivity of the EM procedure to local maxima. An answer to all three questions is provided by a generalization called *annealed EM* [3]. Annealed EM is closely related to a technique known as *deterministic annealing* that has been applied to many clustering problems (e.g. [8, 7]). Since a throughout discussion of annealed EM is beyond the scope of this paper, I will skip the theoretical background and focus on a procedural description. The key idea in deterministic annealing is the introduction of a temperature parameter $T \in \mathbb{R}^+$. Applying the annealing principle to the horizontal hidden variables H the corresponding posterior calculation in (6) is generalized by replacing n_{ij} in the exponent by n_{ij}/T. Downweighting the number of observations

Saul, L.; Jordan, M.I. Learning in Boltzmann trees.
Neural Computation, Nov. 1994, vol.6, (no.6):1174-84.

(a) *Verbatim*
Introduces a large family of Boltzmann machines that can be trained by standard gradient descent. The networks can have one or more layers of hidden units, with tree-like connectivity. We show how to implement a supervised learning algorithm for these Boltzmann machines exactly, without resort to simulated or mean-field annealing. The stochastic averages that yield the gradients in weight space are computed by the technique of decimation. We present results on the problems of N-bit parity and the detection of hidden symmetries.

(b) *Word stems*
introduc larg famili boltzmann machin train standard gradient descent network layer hidden unit connect implem supervis learn algorithm boltzmann machin exactli simul anneal stochast averag yield gradient weight space techniqu present result problem pariti detec hidden symmetri

(c) *Ghost writer*

level 1	paper model base new method gener process differ effect approach provid set studi develop author
level 2	function propos model error method input optim gener neural paramet paper obtain shown appli output
level 3	gener number set neural propos function perform method inform data given obtain approxim dynamic input
level 4	neural pattern rule number process recogni rate perform classif propos gener input neuron time properti data
level 5	converg neural optim method rule rate dynamic process pattern paramet studi statist condition adapt limit
level 6	perceptron exampl error gener rule onlin calcul deriv backpropag simpl output asymptot solution separ unsupervis
level 7	error neural architectur perform entropi statist multilay activ backpropag gener number maximum pattern phase
level 8	teacher delta output introduc sampl replica decai nois projec correl student temperatur gain dynamic predic

Figure 2: (a) Abstract from the generated LEARN document collection, (b) representation in terms of word stems, (c) words with lowest perplexity under the CAM for words not occuring in the abstract (differentiated according to the hierarchy level).

Most frequent words					CAM node top words				
# 33	#35	#42	#50	# 88	# 33	#35	#42	#50	# 88
learn	learn	control	learn	learn	analog	program	feedback	reinforc	interact
network	exampl	learn	algorithm	educ	control	theori	desir	scheme	video
neural	algorithm	robot	network	student	oper	set	dynamic	oper	multimedia
algorithm	gener	model	neural	technologi	implement	space	position	adapt	scienc
weight	problem	propos	propos	develop	backpropag	induc	arm	parallel	remot

Figure 3: Group descriptions for exemplary inner nodes by most frequent words and by the highest probability words from the respective CAM node.

by taking the likelihood contribution to the $1/T$–th power for $T > 1$ emphasizes the prior and will in general increase the entropy of the (annealed) posterior probabilities. In annealed EM, T is utilized as a control parameter which is initialized at a high value and successively lowered until the performance on a held-out set starts to decrease. Annealing is advantageous for model fitting, since it offers a simple and inexpensive way to control the effective model complexity. This avoids overfitting and improves the average solution quality of EM procedures. Moreover, it also offers a way to generate tree topologies, since annealing leads through a sequence of so-called phase transitions. More details on this subject can be found in [3].

4 Results

All documents utilized in the experiments have been preprocessed by word suffix stripping with a word stemmer. A standard stop word list has been utilized to eliminate the most frequent words, in addition very rarely occuring words have also been eliminated. An example abstract and its index term representation is depicted in Figure 2 (a),(b). The experiments reported are some spotlights selected from a much larger number of performance evaluations. They are based on two datasets which form the core of our current prototype system: a collection of 3609 recent papers with 'learning' as a titleword, including all abstracts of papers from *Machine Learning* Vol. 10-28 (LEARN), and a dataset of 1568 recent papers with 'cluster' in the title (CLUSTER).

The first problem we consider is to estimate the probability for a word occurrence in a text based on the statistical model. Figure 2 (c) shows the most probable words from different abstraction levels, which did not occur in the original text of Figure 2 (a). The abstractive organization is very helpful to distinguish layers from trivial suggestions of unspecific word occurrences up to highly specific technical

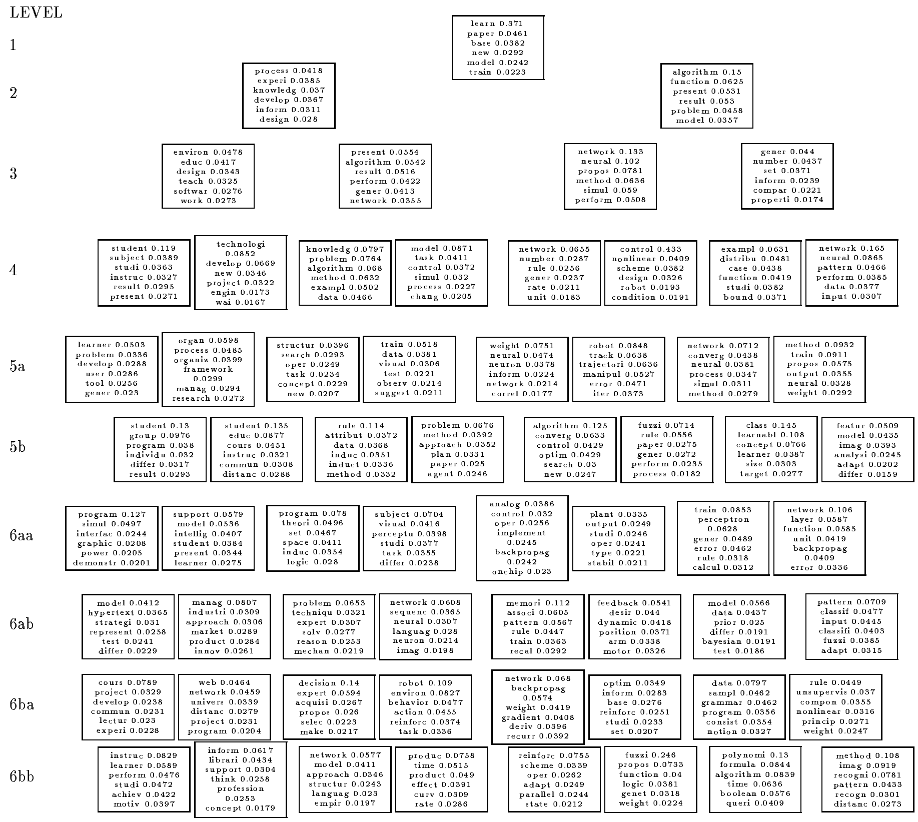

Figure 4: Top 6 levels of the cluster hierarchy for the LEARN dataset. Nodes are represented by their most probable words.

terms.

One of the most important benefits of the CAM is the resolution–specific extraction of characteristic keywords. In Figure 4 we have visualized the top 6 levels for the dataset LEARN. The overall hierarchical organization of the documents is very satisfying, the topological relations between clusters seems to capture important aspects of the inter-document similarities. In contrast to most multi–resolution approaches the distributions at inner nodes of the hierarchy are not obtained by a coarsening procedure which typically performs some sort of averaging over the respective subtree of the hierarchy. The abstraction mechanism in fact leads to a specialization of the inner nodes. This specialization effect makes the probabilities $p_{\bullet|\nu}$ suitable for *cluster summarization*. Notice, how the low–level nodes capture the specific vocabulary of the documents associated with clusters in the subtree below. The specific terms become automatically the most probable words in the component distribution, because higher level nodes account for more general terms. To stress this point we have compared the abstraction result with probability distributions obtained by averaging over the respective subtree. Figure 3 summarizes some exemplary comparisons showing that averaging mostly results in high probabilities for rather unspecific terms, while the CAM node descriptions are highly discriminative. The node–specific word distribution thus offer a principled and very satisfying solution to the problem of finding resolution–specific index terms for document groups as opposed to many circulating ad hoc heuristics to distinguish between typical and topical terms.

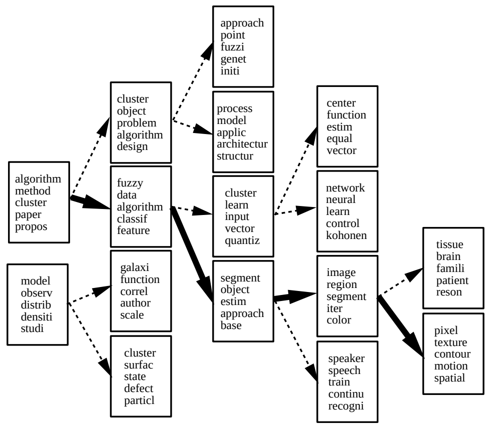

Figure 5: Example run of an interactive image retrieval for documents on 'texture–based image segmentation' with one level look-ahead in the CAM hierarchy.

An example run for an interactive coarse-to-fine retrieval with the CLUSTER collection is depicted in Figure 5, where we pretend to be interested in documents on clustering for texture–based image segmentation. In a real interactive scenario, one would of course display more than just the top 5 words to describe document groups and use a more advanced shifting window approach to represent the actual focus in a large hierarchy. In addition to the description of document groups by inner node word distributions, the CAM also offers the possibility to attach prototypical documents to each of the nodes (the ones with maximal probability $p_{\nu|i}$), to compute most probable documents for a given query, etc. All informations, the cluster summaries by (locally) discriminant keywords, the keyword distributions over nodes, and the automatic selection of prototypical documents are particularly beneficial to support an interactive retrieval process. Due to the abstraction mechanism the cluster summaries are expected to be more comprehensible than descriptions derived by simple averaging. The hierarchy offers a direct way to refine queries and can even be utilized to actively ask the user for additional specifications.

Conclusion: The cluster–abstraction model is a novel statistical approach to natural language learning for information retrieval which has a sound foundation on the likelihood principle. The dual organization of document cluster hierarchies and keyword abstractions makes it a particularly interesting model for interactive retrieval. The experiments carried out on small/medium scale document collections have emphasized some of the most important advantages. Since the model extracts hierarchical structures and supports resolution dependent cluster summarizations, the application to large scale databases seems promising.

References

[1] W.B. Croft. Clustering large files of documents using the single-link method. *Journal of the American Society for Information Science*, 28:341–344, 1977.

[2] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–38, 1977.

[3] T. Hofmann and J. Puzicha. Statistical models for co-occurrence data. Technical Report AI-Memo, to appear, Center for Biological and Computational Learning, Massachusetts Institute of Technology, 1998.

[4] N. Jardine and C.J. van Rijsbergen. The use of hierarchical clustering in information retrieval. *Information Storage and Retrieval*, 7:217–240, 1971.

[5] M.I. Jordan and R.A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2):181–214, 1994.

[6] G.J. McLachlan and K. E. Basford. *Mixture Models.* Marcel Dekker, INC, New York Basel, 1988.

[7] F.C.N. Pereira, N.Z. Tishby, and L. Lee. Distributional clustering of english words. In *Proceedings of the Association for Computational Linguistics*, pages 183–190, 1993.

[8] K. Rose, E. Gurewitz, and G. Fox. Statistical mechanics and phase transitions in clustering. *Physical Review Letters*, 65(8):945–948, 1990.

[9] P. Willett. Recent trends in hierarchical document clustering: a critical review. *Information Processing & Management*, 24(5):577–597, 1988.